Developing Speech Applications for Personal Handheld Devices on TI's OMAP™ Platform

By Dana Myers
Texas Instruments
Marketing Manager,
OMAP Developer Network

TI's OMAP™ platform provides a superior solution for the development of speech applications on personal handheld devices. The power-efficient OMAP architecture combines DSP signal processing for speech with the general-purpose system performance of a RISC processor. An open software architecture is designed to encourage development of speech engines, speech applications and complementary applications such as multimedia. Development support, including a speech recognizer and prototype applications, helps developers create their products quickly and reduces time to market. The OMAP platform ensures that developers can make the most of the growing opportunities for personal handheld devices by adding speech applications easily and flexibly.

Table of contents

Use of speech technology is growing, offering an excellent opportunity for application developers to add valuable functionality to handheld, mobile and wireless personal devices. Today, speech in personal handheld devices is largely limited to voice dialing, but the technology is available for broader development of speech recognition and text-to-speech applications. Developers who are considering adding speech capabilities need to become familiar with the issues of speech technology. These include not only processing and memory requirements, but also how specific platform architectures and support can aid development and reduce time to market.

The potential rewards for those who add value with speech applications are enormous. According to estimates from a variety of market research firms, the compound annual growth rate for personal handheld devices during the next two years is projected to be up to 20 percent, with total device shipments numbering as much as 700 million units in 2004 worldwide. In order to target this huge market with value-added speech applications, developers must look for underlying technology that gives them high performance, power efficiency and support that can help them launch new products quickly.

TI's OMAP™ platform offers a superior solution for the development of speech technology in wireless and other portable devices. The dual-core architecture of the OMAP1510 and OMAP5910 processors incorporates not only a power-efficient digital signal processor (DSP), but also a high-performance RISC core. As a result, both of these OMAP processors deliver the mathematically intensive signal processing needed for speech recognition and text-to-speech, as well as the general-purpose performance needed for system-level operations. The dual-core architecture enables speech applications to operate efficiently, saving power and permitting the system to continue other types of operations. Backed by extensive software development support from TI and TI Third Parties, the OMAP platform presents a powerful solution for developing next-generation, value-added speech applications for personal handheld devices.

Potential speech applications

Speech is a desirable user interface in personal handheld devices, where small displays and keypads make viewing and entry difficult, especially when the user is in motion. As features continue to increase in these devices, navigation of handheld devices could become more complex and difficult to use. Moreover, alternative forms of input and output cannot always be well adapted to personal devices. Navigation devices such as mice are not integral to the handheld system, which by nature must be as compact as possible. Touch screens, though useful at times, are not appropriate for input while in motion.

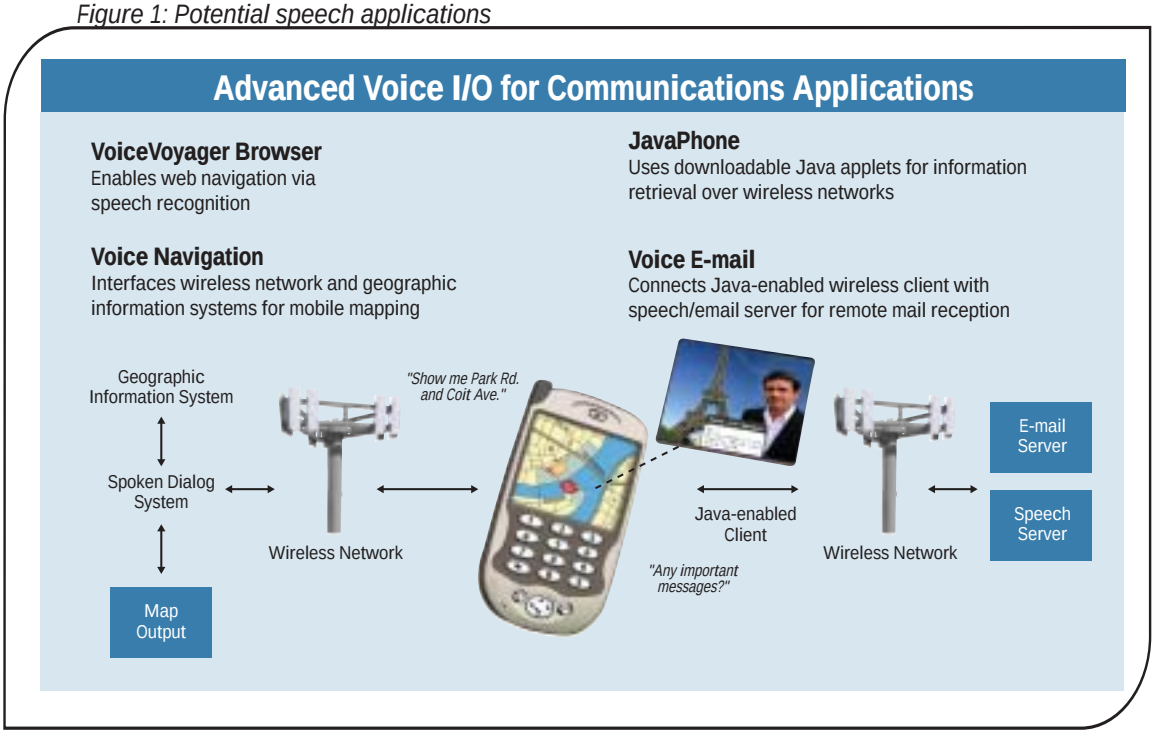
Speech offers a natural input and output modality for human users, and it is safer than other forms of I/O, especially when the user is driving. In most applications, speech is desirable as a complement to keypads and displays, rather than as a replacement. In very noisy environments, for instance, speaking and hearing may be impossible, so the user may have to rely on

keypad entry and display readouts. Similarly, users normally prefer to type items such as PIN numbers and passwords, rather than saying them out loud within earshot of other people.

Voice dialing is the most common application of speech technology in personal wireless devices today. Voice dialing allows hands-free and eyes-free use of the phone to make calls, a feature that is especially useful while driving. Voice dialing can include both name dialing, calling people whose names are on a personal call list, and digit dialing, speaking the digits of a phone number. Other potential speech applications, some of which are shown in Figure 1, include:

- *Voice-enabled email*, which includes surfing the mailbox, dictating email using speech input, and listening to email being read.
- *Information retrieval* of stock quotes, headline news, flight schedules, weather forecasts, and so forth, from the Internet via speech. For example, instead of going to a Web site and typing in one stock ticker at a time or viewing a predefined list, the user could say "My Stock Quotes or Stock Quote, Texas Instruments."
- *Personal information management* which allows the user to make an appointment, check a calendar, add contact information, and so forth, by speaking.
- *Voice browsing* with voice-enabled program menus so that the user can surf Web sites, add speech bookmarks, and listen to items being read back from Web pages.
- *Voice navigation*, a completely voice-I/O-driven system to get directions, for hands-free and eyes-busy conditions.

Figure 1: Potential speech applications



Speech technology issues

In order to suit the needs of personal device users, speech applications must "listen" and potentially be able to "talk". In other words, the applications should be capable of accomplishing both automatic speech recognition (ASR) and text-to-speech (TTS) generation, also known as speech synthesis. Technically, ASR and TTS are different functions, though they share hardware resources and some software modules. From a usage perspective, though, the two functions complement each other in many applications.

Speech systems must satisfy certain baseline requirements for usability. Obviously, speech output must be intelligible so that users can understand it. ASR must also support speech delivered in a natural way, given the purposes of the application. What is considered natural can vary widely, from a small vocabulary and discrete, word-by-word delivery used for names and simple commands, to support for a large vocabulary and normal continuous delivery. People also vary in the natural qualities of their voices, as well as how they pronounce words, so the system should have flexibility to accept the delivery of different speakers. A recognition engine must be accurate or consumers will not use the technology.

Hand-in-hand with the issue of supporting natural speech usage is the question of language support. How many languages is a given system going to recognize? How difficult is reprogramming the system to support a different language, or different sets of languages? For manufacturers dealing with a global market, the answers to these questions can be significant.

Design considerations

- Robustness to noise: Is the system able to function accurately in noisy environments such as automobiles, downtown streets, expressway traffic, airports, and shopping malls, as well as in hands-free uses that move the microphone further away from the speaker's mouth?
- Speaker dependence: Does the system have to be trained by the user, or can anyone use it?
- Continuous versus discrete speech: Can the speaker talk naturally, without pauses between words or digits?
- Flexible versus fixed vocabulary: Can the system recognize any speech, or does it have to be constrained to a particular set of words or tasks?
- Rejection of unintended words: Is the recognizer intelligent enough to reject words that are not in its vocabulary?
- Resource requirements: Can the recognizer fit entirely in a personal device without sacrificing performance? How efficient is the application in MIPS and power consumption?
- Programmability: Is the system able to upgrade its speech algorithms as these evolve, and to modify vocabulary and language support as needed?
- Flexibility: Can the technology adapt to the way the user speaks? Does the platform support complementary high-level applications such as multimedia?
- The need for a speech solution.

Speech technology issues *(continued)*

System requirements for speech are processing-intensive and can include considerable memory, depending on the size of vocabulary supported. To the extent that the application is server-based, wireless bandwidth usage will be increased. These factors affect other system considerations as well. The greater the MIPS and transmission requirements of an application means greater the power consumption on a given system which results in shorter battery life or more frequent recharges. When an application demands memory external to the processor the response time will likely increase.

Some application trade-offs can help limit system requirements by giving up features that may not be necessary for certain uses of handheld devices. A speaker-dependent system that only recognizes discrete speech with a small vocabulary requires considerably fewer resources than a speaker-independent system that recognizes continuous speech with a large vocabulary. Support for additional languages adds to the processing and multiplies the memory required by an application. Features such as noise robustness and barge-in are key, but add to complexity and memory requirements.

Obviously, developers would prefer to compromise the performance of their basic applications as little as possible by adding features such as speaker-independence, continuous speech, vocabulary size and language support. There are options that help minimize the compromise entailed in speech technology, such as distributed speech recognition (DSR). DSR divides the task of recognition, so that the handset converts raw speech into spectral feature vectors, while the server performs the recognition. This approach, and a similar approach for distributed TTS, depends on standardization of processing methods and transmission protocols. Despite the promise of such technologies, however, developers are still faced with limited resources in personal handheld devices for speech applications.

As a result, selecting a platform designed for the needs of high-performance applications such as speech is just as important as tailoring the features of the application. The platform must be processing-intensive yet capable of achieving a high level of power efficiency, not only in core operations but also in handling memory. There should be sufficient MIPS available to support multimedia, security and other complementary applications. Programmability, providing the capability to incorporate new algorithms, is essential. Finally, the platform must include a software architecture designed to support the modular development of applications in order to help developers get their products to market quickly.

OMAP technology: a superior platform for speech

TI's OMAP platform offers a superior solution for the development of speech applications on personal handheld devices. The dual-core architecture of the OMAP1510 and OMAP5910 processors integrates a power-efficient TMS320C55x™ digital signal processor (DSP) and a high-performance ARM9 RISC microprocessor. As a result, these OMAP processors deliver the mathematically intensive signal processing needed for speech, as well as the general-purpose performance needed for system-level operations. The OMAP710 processor is a highly integrated single-chip solution with a TI DSP-based GSM/GPRS baseband for wireless communication processing plus a dedicated TI-enhanced ARM925 processor that power efficiently executes multimedia applications. The OMAP1510, OMAP5910 and OMAP710 processors can support lower-end ARM-based speech applications. They are also code compatible, allowing developers to port software applications to personal products that target different markets. The OMAP1510 and OMAP5910 have the DSP processing power to handle the more intensive speech applications.

As the most flexible platform in the industry for personal devices, OMAP processors support many types of applications that complement speech, including location-based services, security, gaming, personal management, multimedia, biometrics and telematics. TI's Innovator™ Development Kit for the OMAP platform provides a handheld hardware and software system for the development and demonstration of speech applications. All of these support features make the OMAP platform superior for developing speech applications for wireless handheld devices.

Dual-core hardware architecture

The dual-core hardware platform of the OMAP1510 and OMAP5910 is designed to maximize system performance and minimize power consumption. The combination of DSP and RISC cores gives these processors unique advantages in performance and power consumption for personal handheld devices. The RISC is well suited for handling control code, such as the user interface, OS and high-level applications. The DSP, on the other hand, is better suited for the real-time signal-processing required by speech applications.

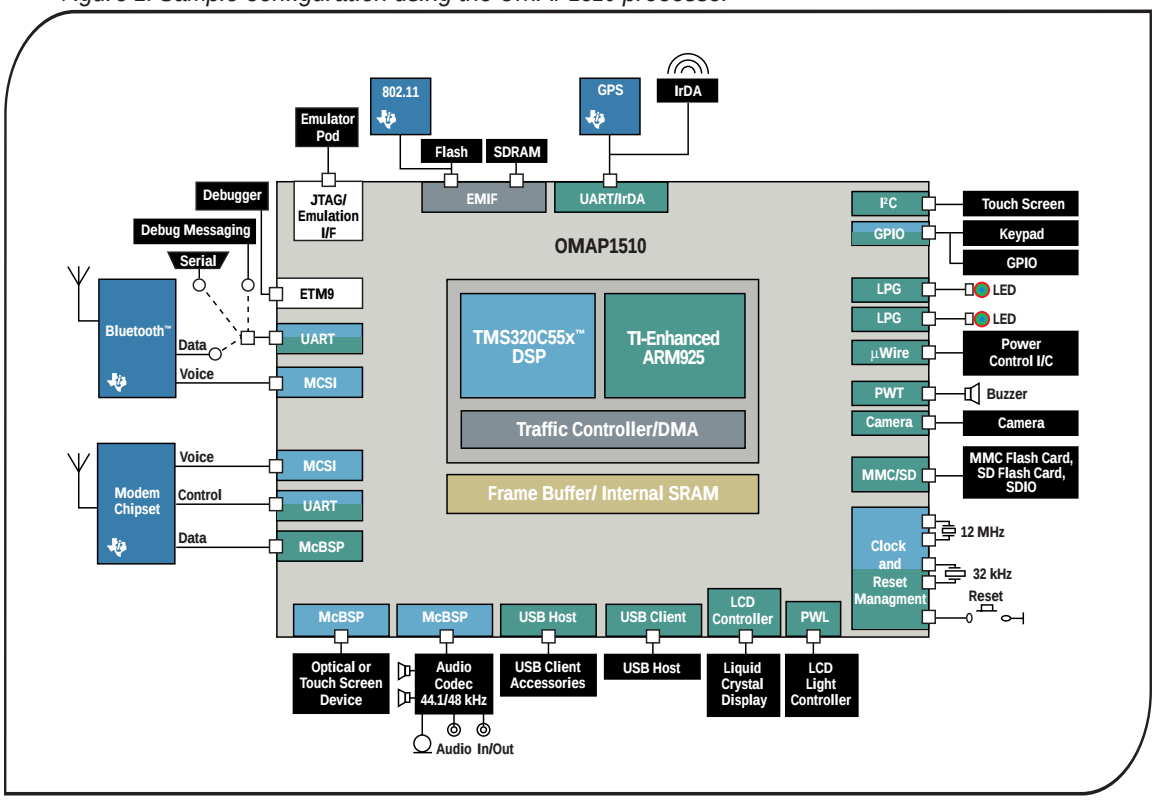
The OMAP1510 architecture, shown in Figure 2, includes on-chip caches for both processors to reduce average fetch times to external memory and eliminate the power consumption of unnecessary external accesses. Memory management units (MMUs) for both cores provide virtual physical memory translation. Low-power operating modes conserve power during periods when the processor is not in use or is used minimally.

The OMAP1510 architecture also contains two external memory interfaces and a single internal memory port. The three memory interfaces are completely independent and allow concurrent access from either core or from the DMA unit. Each processor has its own external peripheral interface that supports both direct connection to peripherals and DMA from the processor's DMA unit. On-chip peripherals such as timers, general-purpose I/O, a UART and watchdog timers support common OS requirements. A color LCD controller is also included.

Dual-core hardware architecture *(continued)*

The OMAP5910 architecture offers system-on-a-chip functionality with peripherals that include 192 Kbytes RAM, USB 1.1 host and client, MMC/SD card interface, multi-channel buffered serial ports, real-time clock, GPIO and UARTs, LCD interface, SPI, uWire and i2s. Like the OMAP1510, the OMAP5910 contains a built-in interprocessor communication mechanism which provides a transparent interface to the DSP for easier code development.

Figure 2: Sample configuration using the OMAP1510 processor



Open software architecture

A key component of the OMAP platform is an open software architecture that supports application development and provides a flexible upgrade capability. This architecture includes a framework for developing software that targets the system design and APIs for executing software on the target system. The architecture supports the standardization and reuse of existing APIs and application software, accelerating time to market for new software products.

To simplify software development, the DSP software architecture is abstracted from the RISC environment. The abstraction is accomplished by defining an architectural interface that allows the RISC to be the system master. The DSP/BIOS™ bridge interface provides communication that enables RISC applications and device drivers to:

- Initiate and control DSP tasks
- Exchange messages with the DSP
- Stream data to and from the DSP
- Perform status inquiries

Designing speech applications for the OMAP platform

Within the OMAP Developer Network TI is working with several key third-party developers who are creating speech technology such as ASR, TTS, DSR and speaker verification. Each of these companies has its own strengths in the marketplace, which they can offer to OMAP customers. At the same time, TI has internally developed speech recognition software designed for small-vocabulary, small-footprint recognition that takes advantage of the dual-core architecture of the OMAP platform. The TI Embedded Speech Recognizer (TIESR) provides:

- Speaker-independent command and control
- Speaker-independent continuous digit recognition
- Speaker-independent continuous speech recognition
- Speaker-dependent name dialing and command and control
- Dynamic grammar and vocabulary capability, which facilitates applications like voice browsing
- Robust performance in noisy environments
- Optional speaker adaptation for enhanced performance

To maximize recognition performance, a unique compensation algorithm handles both additive background noise and the convolutional distortion due to microphone variability. While some developers may prefer to create their own ASR engines, TIESR is available for TI customers who want to use a proven recognizer in their applications.

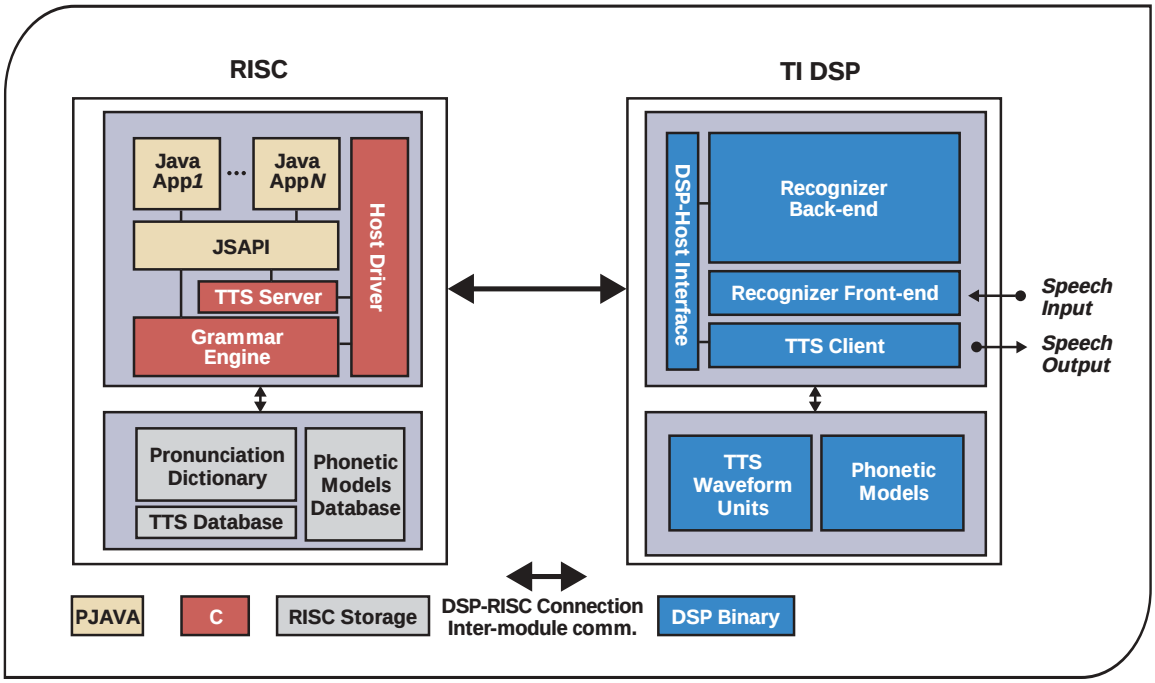
In the TIESR implementation, the computation-intensive, small-footprint recognizer runs on the DSP, while the less computation-intensive, larger-footprint components for grammar, dictionary and acoustic models reside on the RISC. The recognition models are prepared on the RISC and transferred to the DSP via the DSP/BIOS bridge.

Designing speech applications for the OMAP platform (continued)

Figure 3 shows a schematic block diagram of a speech-enabled OMAP application based on a TIESR-like recognition engine. This is only one example of how speech technology can be distributed on the dual-core OMAP processors. The application runs on the RISC, while the entire speech recognizer and portions of the TTS software run on the DSP. The application interacts with the speech recognizer and TTS via a speech API that encapsulates the DSP-RISC communication details.

Two of the most commonly used API standards for the API layer are the Java Speech API (JSAPI) or a variant of Microsoft's Speech API (SAPI). The API is implemented in terms of a set of basic Speech Primitives, which contains functions such as starting and stopping the recognizer and TTS, pausing and resuming audio, loading grammars, setting the TTS rate, and so forth. The hierarchical API architecture makes it easier to incorporate changes in the lower-level API without having to rework the higher-level JSAPI or SAPI. In addition, the Speech Primitives layer reduces the amount of development needed to implement either of the incompatible JSAPI or SAPI standards.

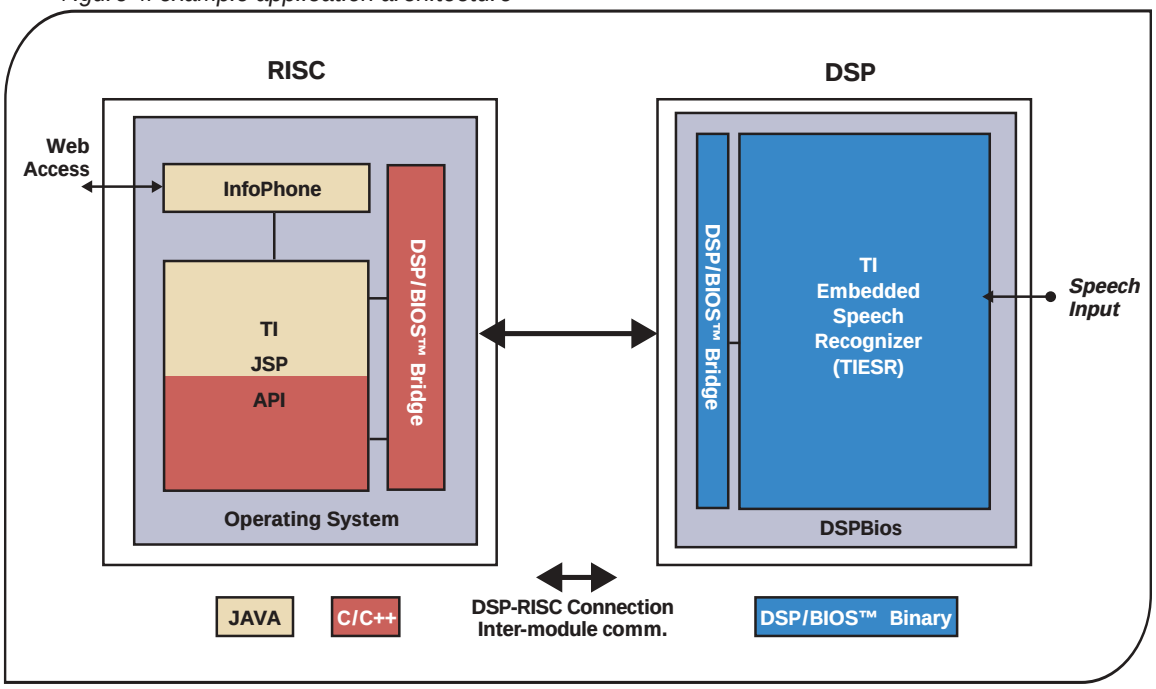
Figure 3: Embedded speech I/O architecture



Speech application example

An example of a speech application based on this embedded architecture is the InfoPhone, developed by TI specifically for the wireless domain. The InfoPhone is a speech-enabled Java application that allows speech-enabled retrieval of useful information. TI has developed three prototype speech-based information services for the InfoPhone to provide users with stock quotations, flight schedules and weather forecasts. Each service includes a vocabulary of up to 50 words, and the system switches seamlessly between the vocabularies because of its dynamic vocabulary capability. The application is designed so that keypad input always stays active during speech, providing flexibility in case of environmental interruptions or if the user needs privacy for input. Figure 4 shows the speech recognition architecture for the InfoPhone example.

Figure 4: example application architecture



Development support

TI's OMAP software and development support help developers bring speech applications to market quickly. Developers have at their disposal for DSP development TI's eXpressDSP™ Real-Time DSP Technology, including the DSP/BIOS real-time operating system (RTOS), the Code Composer Studio™ IDE, and the TI Algorithm Standard that ensures the modular development of off-the-shelf software. The Code Composer Studio for the OMAP platform integrates all host and target tools, including those for the ARM9 RISC core, in a unified environment for easier configuration and optimization. To simplify development even more, the OMAP 5910 and OMAP1510 processor's built-in interprocessor communication mechanism eliminates the need for developers to program the RISC and DSP independently, greatly reducing programming time and complexity.

TI's OMAP Developer network, including developers shown in Figure 5, also provides speech engines and tools for development. To provide flexibility, the OMAP platform supports the most widely used operating systems for personal handheld devices, including:

- Embedded Linux®
- Microsoft® PocketPC2002
- Microsoft® Smartphone2002
- Palm OS®
- Symbian OS™

Figure 5: Third-party speech developers

Figure 6:
The Innovator™ Development Kit for the OMAP™ Platform

In addition, TI has created the Innovator™ Development Kit for the OMAP Platform. As Figure 6 and 7 show, the Innovator development kit provides the hardware and essential software of a personal system to aid in developing speech applications under realistic user conditions.

Figure 7: The Innovator™ Development Kit for the OMAP™ Platform (Deluxe version)

A wealth of opportunity

Speech technology offers a wealth of opportunity to developers of handheld, mobile and wireless personal devices. TI's OMAP platform provides a superior solution for the development of speech applications on personal handheld devices. The OMAP1510 and OMAP5910 processors offer a power-efficient dual-core architecture that combines TMS320C55x DSP signal processing for speech with the general-purpose system performance of an ARM9 RISC processor. OMAP processors include an open software architecture designed to encourage development of speech engines and applications, as well as complementary types of applications such as multimedia.

As the market for personal handheld devices grows, speech applications will increase mobility and functionality for users. The OMAP platform ensures that developers will be able to make the most of the opportunities that this growing market offers.